ELSEVIER

International Journal of Industrial Ergonomics 26 (2000) 217–230

International Journal of
Industrial
Ergonomics

www.elsevier.nl/locate/ergon

Reliability and validity of the Maintenance Resources Management/Technical Operations Questionnaire

James C. Taylor*

School of Engineering, Santa Clara University, 500 EL Camino Real, Santa Clara, California 95053-0590, USA

Abstract

The Maintenance Resource Management–Technical Operations Questionnaire (MRM/TOQ) is a brief survey questionnaire instrument developed to measure the attitudes and intentions of participants in airline maintenance communication and safety training workshops. This present paper describes the test of the effectiveness of the MRM/TOQ for its intended purpose as an evaluative measure. The test included samples of both maintenance management and aviation maintenance technicians (AMT) from the same airline who attended similar workshops several years apart. The test of the MRM/TOQ reveals an integrity of factor structure and a reasonable reliability of both individual items and multi-item scales. Tests of concurrent validity indicate that many of the items and scales measure what they purport to measure. Testing construct validity is important, but is often difficult. The various measures of fixed response items and open-ended questions in the questionnaire allowed a test of construct validity which indicated that what respondents reported about the importance of several topics was consistent with predicted responses about what would be implemented from the training. The MRM/TOQ has reasonable psychometric characteristics, and it can be used as a robust measure to evaluate communication and MRM training programs in other aviation maintenance settings.

Relevance to industry

Many airlines use surveys to understand the attitudes, opinions, and perceptions of their employees. It is important that surveys are scientifically "good" measures. One survey questionnaire, the "MRM/TOQ", has become widely used for assessing communication and management improvement programs in aviation maintenance. Evaluating the reliability and validity of the MRM/TOQ helps confirm its scientific "goodness" as a measurement instrument.

Keywords: Construct; MRM; Psychometric; Questionnaire; Scale; Statistic; Survey

1. Introduction

Interest in resource management for maintainers and aviation technical operations has grown rapidly over the past decade. "Maintenance Resource Management" (MRM) training, (also known as "Maintenance Human Factors training") is a response to that interest. MRM training has been evaluated, in part, by attitude survey questionnaires, which have been examined using FAA and NASA sponsorship. The measuring instrument used in the evaluation of MRM training is called

* Corresponding author. Tel.: +1-408-554-4154; fax: +1-408-554-5474.

E-mail address: jctaylor@scu.edu (J.C. Taylor).

0169-8141/00/$ - see front matter
PII: S0169-8141(99)00068-2

the Maintenance Resource Management Technical Operations Questionnaire (MRM/TOQ). Earlier this questionnaire was called the Crew Resource Management Technical Operations Questionnaire, or "CRM/TOQ" (Taylor and Robertson, 1995). The instrument has become more widely used and known as the "MRM/TOQ". Our use of that name here acknowledges the current industry term (cf., Taylor, 2000). The MRM/TOQ includes attitude measures which were modified and developed specifically for use in aviation technical operations. This paper describes that questionnaire and presents the structure of the attitude measures with respect to MRM issues as well as the "goodness" of those measures for program evaluation. This test of psychometric characteristics of the MRM/TOQ includes the responses from aviation maintenance technicians (AMTs) as well as from other occupations employed in airline technical operations.

1.1. The challenge of testing the goodness of psychometric measures

When a specific, observable event is selected as an indicator of a general, unobservable concept, one is faced with the phenomenon of psychological inference. Because such inferences can vary widely, it is important to have an understanding of the degree to which that the events fit the concepts in question. Questionnaire survey items were chosen by an airline as observable indicators (measures of events) to represent the otherwise unobservable attitudes and perceptions of their Technical Operations personnel. These items, or measures, embodied operational definitions of concepts included in the company's MRM training – such as communication, self-knowledge, situation awareness, goal attainment and personal assertiveness. How well do the operational definitions and their resulting observations represent those variously abstract concepts? That degree of representation can be considered the goodness of the measures.

Empirical tests of reliability and validity are used to examine that goodness. A test of validity seeks to determine if the measure used is truly a measure of the concept as operationally defined. The related issue of reliability asks whether the measure used can be depended upon to yield the same value in repeated independent assessments of the same participants. The two concerns differ from one another, because a reliable measure can be invalid, and a valid measure can be erratic. They are however, not completely independent of one another since the reliability of a measure places a ceiling on its observed validity – in other words a measure can never be more valid than it is reliable.

The potential multiplicity of operational definitions of a concept is an important matter for both reliability and validity. On one hand multiple measures can create disparate observations which will correlate differently with different outcome or end result variables. On the other hand multiple measures can also converge to render a joint observation which better represents the rich complexity of the real world. In obtaining a "good" measure of the complex concepts of MRM training an important goal is to use a survey, proven in other (but similar) occupations, that contains enough items to provide convergence to a smaller set of concepts while not requiring undue time or effort by the participants.

Three aspects of "goodness" of the MRM/TOQ are examined here. The first is the structure of the questionnaire, the second is reliability of items and scales, and the third is validity of the scales. The method used and results of the examination of those aspects will be presented after a brief description of background information.

1.2. Prior experience in measurement of attitudes in the cockpit

Gregorich et al. (1990) published evidence that their Cockpit Management Attitudes Questionnaire (CMAQ) is useful as a training, evaluation and research tool. Their questionnaire contains 25 items chosen to measure attitudes that are either conceptually or empirically related to CRM. Their study involved the further enhancement of their instrument by revealing a consistent internal structure, which permitted the calculation of several scales or indices within the overall questionnaire. Such scales permit more specific scores than a single total index and provide more accurate and reliable results than are available from each of the individual questionnaire items alone. Gregorich

et al. divided the CMAQ results into several relevant subscores using Factor Analysis which uses relationships among the survey respondents' own answers to the individual questions. When applied in this way the subscales, or "factors", that result from the analysis can be used for greater ease in understanding the outcomes, as well as improvement in the strength of the measurement. A similar strategy was undertaken for the data obtained with the MRM/TOQ in the present study.

2. The data

The Technical Operations division of one airline provided the data for this study. From that population, two samples were used in this study – the mechanics and inspectors (together, considered AMTs) formed one sample and all other technical operations support and management personnel formed the other sample. That US air carrier conducted its MRM training program between 1991 and 1994. This training was developed internally by the airline and was initiated full-scale in June 1991. During the next 24 months the training was given to technical operations support and management personnel. That first phase involved training over 2000 directors, managers, supervisors, assistant supervisors, engineers, planners, trainers, coordinators and administrative assistants from maintenance, planning, engineering, and quality departments. The training was then expanded to include AMTs, and some 190 mechanics and inspectors participated between September 1993 and March 1994.

2.1. Measurement and training at the company

With the assistance of investigators at the University of Southern California, the company developed a brief survey questionnaire – the MRM/TOQ. This questionnaire was based on previous surveys, one of which (CMAQ) had been used to evaluate cockpit resource management training for flightcrews in that airline; and another survey which had been developed for maintenance personnel in another airline (Taggart, 1990). The resulting MRM/TOQ is a self-report measure of attitudes that are related (conceptually or empirically) to resource management in maintenance and maintenance support functions. The questionnaire asks training participants to express their degree of agreement with a series of statements on a five-point scale where 1 signifies low agreement and 5 signifies high agreement. It was designed to measure attitudes, opinions and perceptions about behaviors and skills the upcoming MRM training was intended to influence – thus establishing a pre-training baseline for gauging the post-training results.

The basic MRM/TOQ contains 26 multiple response questions. The company's changes to the CMAQ involved removing five attitude questions from the original 25, and adding six other questions. The five questions removed, either lacked predictive validity (as reported by earlier flight crew studies; Helmreich et al., 1986), or they lacked relevance to the work of technical operations personnel. The six questions added to the questionnaire measured respondent perception of behaviors dealing with attainment of work goals in the technical operations division. The work of technical operations personnel is seen to differ from flight crew members in the typically longer time required for technical operation's goal attainment and the relatively greater variety of those goals. In addition to the 20 attitude items and 6 goal-attainment items some questions about the respondent's background (job title, city, department, age, experience, and gender) were included.

2.2. Baseline attitude questionnaires

During the late Spring of 1991, the basic questionnaire and an accompanying letter from the Technical Operations Senior Vice President was sent to the homes of all Tech Ops directors, managers, supervisors, and assistant supervisors (1787 total). Within two months 900 completed questionnaires had been received at the University for a return rate of just over 50%. This questionnaire administration preceded the first MRM/TO training class by one week. It was termed the "baseline" survey.

A similar baseline questionnaire and cover letter was mailed to the homes of a random sample of

500 AMTs during the Summer of 1993. By late Autumn 290 questionnaires had been returned (58% return rate). AMTs began participating in MRM training in September 1993.

2.3. Training evaluation questionnaires

An identical questionnaire to the baseline instrument, called the "pre-training" survey, was administered at the beginning of each training session.

A third questionnaire was completed at the end of the training session. This third questionnaire, called the "post training" survey, contained all 26 fixed-response items of the "baseline" and "pre-training" surveys, plus four fixed response items (here numbered 27–30) which provide evaluation of the course itself. In addition the "post seminar" questionnaire includes three open-ended questions (numbered 31–33) requiring written responses.

3. The structure of MRM attitudes

3.1. Factor analysis

Among other statistical procedures undertaken to assess the goodness of the MRM/TO questionnaire measure, Factor Analyses were performed, using "baseline" questionnaire data separately for the 900 management/staff and the 290 AMTs. In both cases a first Factor Analysis was run to determine the underlying internal structure among the answers to the 20 attitude questions in the MRM/TOQ. A second Factor Analysis was undertaken for both management and AMT samples in order to examine the underlying structure of the six goal attainment items for both groups.

Several tests were conducted (Bartlett's test of sphericity, $p < 0.000$; an anti-image correlation matrix, $r < 0.09 < 15\%$; and the Kaiser–Meyer–Olkin measure > 0.74) to establish the appropriateness of these data for Factor Analysis (Norusis, 1990, pp. B127–B131). For each of the two analyses for each of the two samples a principal components analysis was run and initial factors were extracted based on Eigenvalues. Then, from the scree plot the appropriate numbers of the factors were determined as specified by Norusis (1990). Orthogonal varimax rotation was used, employing the final rotated factor matrix. In all cases the factor solution offered good parsimony, interpretability and simple structures.

The results of the Factor Analyses of the 20 attitude questions for the management/staff and AMT samples respectively are presented in Tables 1 and 2. The structures of the MRM/TOQ attitude data (i.e., from the first 20 questions) for both the management/staff sample and the AMT sample were found to resemble the structure reported by Gregorich et al. for their CMAQ. On the strength of this similarity between studies it was decided to apply, where possible, the same names to the MRM/TOQ factors as those developed from the CMAQ. The factors (with the descriptive labels) and their constituent items, are shown below.

Comparing the factor structure of the two samples shows that the differences are primarily ones of relative emphasis and not of substance. Both samples show command responsibility and communication and coordination as first and second factors, and showed avoiding conflict as the third factor. The management sample shows two items from the AMT communication and coordination factor loading on a fourth factor, but both samples show recognition of stressor effects as the next factor. Tables 1 and 2 confirm that the composition of four factors in both analyses is quite similar. The factors derived in these present analyses are similar to that found in a large flight crew sample examined by Gregorich et al. (1990), except that the third factor in precedence in our Technical Operations samples, avoiding conflict, is the fourth factor for the flight crews. The last factor in each Tech Op sample is not used further. That last factor differs for the two samples, it is not reported in Gregorich et al., and has low eigenvalues and accounts for little variance in the total structure in both samples. The item scores in the four factors (except management factor IV) were subsequently averaged for each respondent, to become four new index scores (Taylor and Robertson, 1995; Taylor et al., 1997).

An additional Factor Analysis was undertaken for the six questions added to MRM/TOQ which measured respondent perception of behaviors dealing with attainment of work goals in the technical operations division. Since the 20 items previously

Table 1
Factors loadings: technical operations management and support staff ($n = 900$)

		Factor I	Factor II	Factor III	Factor IV	Factor V
Factor I:	Communication and Coordination					
	18. Manager responsibility is coordination between groups	0.66				0.14
	16. Pre-assignment briefing is important	0.62	0.13		0.21	
	12. Debriefing/critique is important for coordination	0.58			0.19	
	10. Managers should encourage questions	0.56	− 0.18			
Factor II:	Command Responsibility					
	11. Subordinates should not take control		0.69			
	6. Managers should take charge in emergency	0.10	0.63			0.17
	8. Subordinates should not question managers decisions	0.22	0.56	0.15		
	13. Technical proficiency causes successful management		0.47	0.38		
Factor III:	Avoiding Conflict					
	2. It is important to avoid negative comments about others			0.82		
	1. We should avoid disagreeing with others		0.23	0.76		
Factor IV:	Support of Others					
	5. We should be sensitive to other's problems	0.11			0.67	
	14. Training is a most important mgt responsibility.	− 0.21	0.16		0.52	
	17. Coordination requires taking other's personalities into account	− 0.22			0.66	0.10
Factor V:	Recognition of Stressor Effects					
	9. I perform effectively even when fatigued	− 0.15		0.11	0.11	0.73
	20. My decision making is good in abnormal situations	0.19	0.26			0.68
	19. Managers can leave personal problems behind	0.29	0.35	0.11	− 0.20	0.43
	3. Casual conversation improves coordination		− 0.10	− 0.12	0.22	0.21
	4. Good communication and coordination are important for safety				0.16	− 0.13
	7. Management should inform us of plans and actions	0.28			− 0.20	
	15. Coordination is more important in emergencies	0.23	0.21		0.25	− 0.28
Eigenvalues =		2.65	2.28	1.38	1.16	1.05
Percent Variance Accounted for		13%	11%	7%	6%	5%

described measured attitudes or feelings of the respondents, it was felt that the behavioral focus was different enough to require a separate analysis of these six items. The resulting combination of items formed two factors for the management sample while the six items showed less differentiation for the AMT sample. Table 3 shows the derived factor structure for both samples.

Mean scores of constituent items were calculated for each management respondent to form two additional index scores, while all six items were combined for those respondents in the AMT sample.

Table 2
Factor loadings: technical operations aviation maintenance technicians ($n = 290$)

		Factor I	Factor II	Factor III	Factor IV	Factor V
Factor I:	Communication and Coordination					
	18. Manager responsibility is coordination between groups	0.78				0.17
	16. Pre-assignment briefing is important	0.75		0.10		
	15. Coordination is more important in emergencies	0.73			0.10	
	4. Good communication and coordination are important for safety	0.59	− 0.12		− 0.16	0.13
	14. Training is a most important management responsibility	0.57		0.11		− 0.33
	12. Debriefing/critique is important for coordination	0.57				0.43
	17. Coordination requires taking other's personalities into account	0.56				0.25
Factor II:	Command Responsibility					
	11. Subordinates should not take control		0.67	0.18	0.10	
	13. Technical proficiency causes successful management		0.66		0.15	
	6. Managers should take charge in emergency	− 0.13	0.65	0.15	− 0.12	
	8. We should not question superiors actions		0.43	0.46	0.18	
Factor III:	Avoiding Conflict					
	2. It is important to avoid negative comments about others		0.18	0.77		0.18
	1. We should avoid disagreeing with others		0.20	0.75		
	3. Casual conversation improves coordination		− 0.12	0.52		0.10
Factor IV:	Recognition of Stressor Effects					
	20. My decision making is good in abnormal situations			− 0.10	0.74	
	19. Managers can leave personal problems behind		0.19		0.74	− 0.11
	9. I perform effectively even when fatigued		− 0.31	0.29	0.57	0.14
Factor V:	10. Managers should encourage questions	0.11		0.16		0.79
	5. We should be sensitive of others problems	0.28	− 0.39		0.18	0.36
	7. Managers should inform us of plans and actions	0.38				0.46
Eigenvalues =		3.63	2.42	1.50	1.42	1.14
Percent Variance Accounted for		18%	12%	7.50%	7%	6%

4. Reliability of the MRM/TOQ item and index measures

Whenever a survey measure is used, there are many elements which enter into the error component of an individual's answer or score. Some of these include the individual's misreading of an item, distractions, daily fluctuations in an individual's health and emotional status, or the sampling of items or measures used. The size of this error component is related to the reliability of any questionnaire measurement. The smaller the error component or error score, the more reliable the measure is. The term reliability may refer to either of two meanings. *Stability* is a form of reliability which denotes the capacity of a scale to produce

Table 3
Factor loadings

	Factor A	Factor B
Technical Operations Management and Support Staff ($n = 900$)		
21. I am informed of the goals of this organization	0.82	0.15
22. We understand and agree on work goals	0.81	0.22
23. My subordinates receive feedback on our performance	0.72	0.17
24. My subordinates can voice concerns about goals	0.65	0.13
26. Other groups act as if they share our goals	0.23	0.85
25. Other groups plan and coordinate with my subordinates	0.16	0.89
Eigenvalues =	2.90	1.07
Percent Variance Accounted for =	48%	18%
Technical Operations Aviation Maintenance Technicians ($n = 290$)		
23. My crew receives feedback on our performance	0.83	
22. We understand and agree on work goals	0.82	
21. I am informed of the goals of this organization	0.80	
26. Other groups act as if they share our goals	0.75	
25. Other groups plan and coordinate with us	0.73	
24. My subordinates can voice concerns about goals	0.66	
Eigenvalues =	3.52	
Percent Variance Accounted for =	59%	

similar or the same measurement values over time in circumstances when expected. *Consistency* is a second meaning of reliability. Consistency is concerned with the equivalence of an individual's position on different scales intended to measure the same characteristic. Consistency reliability permits the researcher to continue to infer that the same meaning is being used by the respondent and that the item is being used to measure the same thing. A reliable questionnaire leads to measurement results which are similar from time to time, if nothing else changes.

A variety of reliability coefficients for questionnaire items, and multi-item indexes were computed in this present study of AMT and Technical Operations managers' and supervisors' attitudes about interpersonal behaviors and skills, management of stress, problem solving/decision making, and goal attainment. The various types of reliability produced are described below.

4.1. Use of communalities from the Factor Analysis

According to Guilford (1936, p. 477) item communalities (k^2), derived from Factor Analysis may be used as lower bound estimates of item reliability (a measure of internal consistency). Item communalities for the MRM/TOQ are available from the four Factor Analyses described above. The communalities (k^2) for each MRM/TOQ item which is combined into one of the six factor indices for the management group and AMT sample are presented in Tables 4 and 5.

Although the communalities of the items in Factor I and II are lower than expected for both samples, those for items in the remaining factors are moderately high. These data indicate a *lower bound* estimate of intra-item consistencies which are reasonable, if not excellent. More information on the internal consistency of the factor scores, or indices is available using Cronbach's *Alpha*.

4.2. Use of Coefficient Alpha for estimating index reliability

Coefficient Alpha has long been used as a gauge of internal consistency for use in psychometric test development (Cronbach, 1951). As such it is routinely applied as a reliability measure for multiple item indices, and has been shown to be a generalized version of other widely used reliability coefficients measuring internal consistency. Coefficient Alpha can be interpreted as a measure of the reliability of a composite measure when a combination of items is expected to have a large common factor. It also indicates the homogeneity of an item group.

For both the management and AMT samples, the reliabilities using coefficient Alpha were good for scales of this length. Alpha coefficients ranged between 0.54 and 0.56 for Factors I, II, III for the management sample and between 0.51 and 0.77 for the AMTs. The scores for these first three factors compare quite closely with their counterparts reported in Gregorich et al. (1990). An Alpha of 0.39 for Factor V for the management sample was lower than desired for confidence as a stable index. We should note, however, the communalities for two of the component items in this index (Table 4) were reasonably high at 0.58 and 0.59. The Alpha of 0.48

Table 4
Item communalities of maintenance attitude index components aviation maintenance managers and staff support personnel ($n = 900$)

Item		k^2
Factor I:	Communication and Coordination	
	10. Managers should encourage questions	0.41
	12. Debriefing is important for coordination	0.44
	16. Pre-assignment briefing is effective	0.43
	18. Mgr responsibility is coordination bet gps	0.47
Factor II:	Command Responsibility	
	6. Mgr. takes charge in emergency	0.48
	8. Subs do not question mgr's decisions	0.38
	11. Subs do not take control	0.36
	13. Technical proficiency is successful mgt	0.39
Factor III:	Avoiding Conflict	
	1. Avoid disagreeing with others	0.64
	2. Avoid negative comments re. others	0.60
Factor IV:	Support of Other's	
	5. Sensitive to others problems	0.49
	14. Training is mgt. responsibility	0.34
	17. Take other's personalities into account	0.39
Factor V:	Recognition of Stressor Effects	
	9. I perform to other's problems	0.59
	20. My decision making is good in crisis	0.58
	19. Mgr. can leave personal problems behind	0.39
Factor A:	Goal Attainment with my group	
	21. Informed of the goals of this org.	0.69
	22. Managers agree on work goals	0.71
	23. My subs get feedback on our perform	0.54
	24. My subs voice concerns about goals	0.43
Factor B:	Goal Attainment with other groups	
	25. Other groups coordinate with my subs	0.78
	26. Other groups share our goals	0.81

Table 5
Item communalities of maintenance attitude index components aviation maintenance technicians (AMTs) ($n = 290$)

Item		k^2
Factor I:	Communication and Coordination	
	12. Debriefing is important for coordination	0.52
	14. Training is mgt. responsibility	0.45
	16. Pre-assignment briefing is effective	0.59
	17. Take other's personalities into account	0.39
	4. Good comm and coord are important	0.40
	15. Coord more important in emergency	0.56
	18. Leader is responsible for coordination	0.65
Factor II:	Command Responsibility	
	6. Mgr. takes charge in emergency	0.49
	11. Subs do not take control	0.50
	13. Technical proficiency is successful mgt	0.49
	8. Subs do not question Mgr's decision	0.43
Factor III:	Avoiding Conflict	
	1. Avoid disagreeing with others	0.61
	2. Avoid negative comments re. others	0.66
	3. Casual conversation improves coord	0.30
Factor IV:	Recognition of Stressor Effects	
	9. I perform well when fatigued	0.53
	19. Mgr. can leave personal problems behind	0.60
	20. My decision making is good in crisis	0.57
Factor A:	Goal Attainment with others	
	23. My gp gets feedback on our perform	0.83
	22. Managers agree on work goals	0.82
	21. Informed of the goals of this org	0.80
	26. Other groups share our goals	0.75
	25. Other groups coordinate with my gp	0.73
	24. My gp voices concerns about goals	0.66

for the AMTs factor IV is slightly higher, and the communalities between 0.53 and 0.60 are reasonable as well.

The Alpha coefficients for the two management goal attainment factors A and B were 0.77 and 0.74, respectively. These results represent very good internal consistency for these measures of perceived behavior. The single AMT goal attainment factor has an Alpha coefficient of 0.86.

4.3. Test–retest method of reliability computation

With the test–retest technique, the questionnaire should be administered and then at a later time readminister the same test was to the same individuals. A correlation coefficient is then computed between the two sets of scores.

In the present case, the MRM/TO "baseline" questionnaire and the MRM/TO "pre-seminar" questionnaire (virtually the same instrument) were completed by a small sample of management participants within a five-week period, and were identified by the respondents' using the individual code numbers they noted on their questionnaires. The baseline questionnaires were sent to respondents' homes (during the week of 5/26/91) and completed before they arrived, as participants, for the training. These same participant/respondents completed the pre-seminar questionnaires in the training room at the onset of the session (either 6/26/91 or 7/9/91). A total of 13 participants completed both questionnaires, and thus provide the data for calculating test–retest reliability coefficients (using Spearman Rho Correlation Coefficient test for the statistic) for the first 26 items in all versions of the MRM/TO questionnaires. The indices composed from items loading on the four factors in Factor Analysis for attitudes, and the two factors in Factor Analysis for goal attainment were calculated for the individual respondents in this test–retest sample as well. These index scores permitted the computation of the Spearman Rho coefficients for the combined measures over time. Table 6 presents the test–retest reliabilities for the 26 items and six indices of combined items. It can be seen that the vast majority of items have high coefficients of stability. In fact only two of the 26 individual items had test–retest coefficients less than 0.74. When individual items are combined into the six factors derived as described above, the index reliabilities increase further from a low of 0.80 to a perfect coefficient of 1.0 for this sample of 13 respondents. These test–retest coefficients are shown in Table 6.

Table 6
Test–retest method of reliability

Items	ρ
1. Avoid disagreeing with others	0.99
2. Avoid negative comments about others	0.77
3. Conversation increases coordination	0.85
4. Communication important as tech proficiency	0.74
5. Sensitive to others' problems	0.89
6. Manager should take charge in emergency	0.76
7. Manager verbalizes plans	0.57*
8. Subordinates should not question mgr's decisions	0.99
9. I perfom well even when fatigued	1.00
10. Managers encourage questions	1.00
11. Subordinates should not take control anytime	1.00
12. Debriefing is important for coordination	0.87
13. Tech proficiency equals successful management	0.80
14. Training is a management responsibility	1.00
15. Team coordination is important in emergency	1.00
16. Pre-assignment briefings are effective	1.00
17. Should take other's personalities into account	0.87
18. Manager should coordinate between work groups	1.00
19. Manager can leave personal problems behind	1.00
20. My decision making is good even in crisis	1.00
21. Kept informed about our goals/objectives	0.82
22. We understand and agree on work goals	0.97
23. My subs receive feedback on our performance	0.90
24. Subordinates can voice concerns about goals	0.89
25. Other groups plan and coordinate with my subs	0.95
26. Other groups share our goals	0.55*
I. Command responsibility	0.99
II. Communication and Coordination	0.95
III. Avoiding Conflict	0.89
IV. Recognition of Stressor Effects	1.00
A. Goal Attainment with group	0.95
B. Goal Attainment with other groups	0.80

Note: ρ = Spearman rho correlation coefficient ($n = 13$)
*Significance level = 0.05; (all other correlations significant at 0.01 or higher).

4.4. How reliable is the MRM/TOQ?

Based on the three tests above we can conclude reasonable reliability, in both stability and consistency, for most of the 26 items in the MRM/TOQ. When individual items are combined into the six indices suggested by the results of the Factor Analysis, the joint reliabilities further improve. Two parallel tests of managers and AMTs over a 30 month period confirm this reliability of the measures and confirm each sample's performance on the scales tested.

Additional evidence for the stability of the measures over time is reported in Taylor and Robertson (1995). Over 2000 maintenance managers completed the MRM/TOQ before and immediately after the training and, again, 2 months, 6 months and one year following their training. After statistically significant post training increases in average scores, the communication and coordination index, the command responsibility index, and the stress recognition index remained stable in

all three subsequent surveys. The assertiveness index did not increase immediately after the training, but it did increase significantly in the two month follow-up survey, and it remained stable thereafter.

5. Validity of multiple-response questions

The essential question in assessing validity asks: Does the MRM/TOQ survey measure those things the MRM training is intended to change? The objective of investigating the validity of the MRM/TOQ survey is to confirm that attitude data collected are appropriate to the purpose of evaluating the effects of MRM training. The six survey items measuring perception of goal attainment were not expected to change immediately following the training – and this was subsequently established (cf., Taylor and Robertson, 1995), so they are not included as a part of the present test of item validity.

Both *statistical or concurrent validity* and *construct validity* will be demonstrated for the MRM/TOQ. The former type of validity, "concurrent validity", correlates the attitude items and composite scales from the MRM/TOQ with other items which are intended to measure similar aspects of the MRM training. Following Cronbach and Meehl (1955), the application of construct validity in the present study views the MRM/TOQ items as part of a network of related concepts (the "construct"), in this case representing (and leading to) usability of the training material.

5.1. Criterion measures

Concurrent measures of those 20 attitudes already described (cf., Tables 1 and 2) were collected immediately following the training in the "post training" survey. The responses of 353 training participants were included in an analysis of statistical validity using concurrent measures. The sample of 353 included 163 managers who attended the first eight training sessions between June and August 1991 and all 190 AMTs who went through the training between September 1993 and March 1994. These measures were coded to match several objectives of the training. The measures involved either *recognition* of concepts taught in the course (by participants' responses to fixed stem, multiple response questions, numbered 27–30) or respondents' *recall* of those concepts as measured by their written responses to open-ended questions (numbered 31–33).

5.2. Statistical validity

Validity of the basic 20 multiple response "attitude" questions in the MRM/TOQ was empirically tested by correlating the attitude scores with the scores of participants responses of training experience and evaluation. This evidence of training value or "usefulness" was obtained from a set of fixed-response items, a subset of question 27. For this test, these judgments of the participants regarding the usefulness or value of the training were considered to relate to the same concepts as their attitudes expressed in the 20 core attitude measures. The same test was applied to the four composite indices constructed from the results of the Factor Analysis of those 20 items.

Evidence for this criterion-related statistical validity (sometimes called "concurrent validty", cf., Selltiz et al. (1976)) consists of a statistically significant relationships between the 20 survey measures and the four indices, with the several fixed response criterion measures (Q 27) of the course itself. The criteria are established by comparing training course objectives with the attitudes measured in the MRM/TOQ. Those objectives of the course were presented by the trainers to participants during the introduction to each two-day session. These course objectives themselves remained stable. They were phrased as helping participants to

1. enhance interpersonal skills,
2. promote assertive behavior,
3. understand and manage stress,
4. understand individual leadership styles,
5. enhance rational problem solving and decision making skills,
6. diagnose "norms" and their effect on safety.

Objectives "1" through "5" can be measured in the classroom and following training by asking students for demonstration and judgment. The sixth objective "Diagnose norms *and their effect on safety*" (emphasis added) is virtually impossible to test against classroom criteria. Although one might

be able to measure comprehension of the concept of safety norms, safety data would be required to empirically validate safety effects of learning about norms.

Question 27 asked the participants, "How do you rate the following aspects of the training?" This stem was followed by a list of seven to nine aspects of the course. The number of aspects included depended upon several changes in this evaluation question which were made between the beginning of the training in 1991 and its conclusion in 1993–94. These changes were intended to reflect improvements in the course curriculum and to make the questionnaire more appropriate to the course as it evolved. There were five course aspects in question 27 which remained unchanged between 1991 and 1994 and can thus provide comparison criterion measures for both management and AMT samples.

The comparable concepts in the question 27 evaluation item are as follows

(Rated from: 1 = "waste of time", to 5 = "extremely useful")

(1) interpersonal communication,
(2) assertiveness,
(3) stress effects and stress management,
(4) analysis of personal styles,
(5) problem solving skills using role playing.

The rated scores for each of these five criteria were correlated with the 20 attitude items and four composite indices which were intended to measure the same aspects of the training. Such correlations indicate concurrent validity of the items and index scales. Table 7 presents these validity coefficients.

The responses to the fixed response questions of training value (Q27) in the post-training MRM/TOQ correspond reasonably well with the first five course objectives. Table 7 shows that over 40% (22) of the 54 correlations for the 20 individual attitude items were significant above the 0.05 level of significance (one tail test) and were in the expected direction. By chance alone, only two (and a fraction) of these 54 validity coefficients would be expected to meet or exceed the size of the 22 correlations reported in Table 7. For the composite measures 58% (7) of the 12 validity coefficients expected were significant. Although the statistically significant correlation coefficients in this analysis are not large in size they are consistent in direction and numerous when compared to the number expected by chance alone.

These statistical validity data clearly show that two of the four multiple item scales (Index I and II) correspond to the appropriate topics and course modules. The assertiveness course topic did not relate inversely with Index III, "Avoiding Conflict". The stress management topic was also not found related to any of the individual items expected to measure stress effects, or to the composite Index IV, "Recognition of Stressor Effects".

5.3. Construct validity of the criterion measures

This type of validity (cf. Cronbach and Meehl, 1955) views the MRM/TOQ items as part of a network of related concepts representing usability (or intended implementation) of the content of the training material – on a theoretical continuum from quick and easy for a person to apply at one extreme; to time-consuming, difficult to master, or requiring the close cooperation of others at the other extreme.

In the post-training version of the MRM/TOQ the participants were presented with a list (question 27), previously described as criterion measures, of specific topic areas from the course and asked to rate them for their value on a five-point scale. The post-training questionnaire also requested participants to write-in which parts of the course they thought were "particularly good" (question 31), and how they intended to "use the training" when they returned to their work place (question 32). If the fixed response items (Q 27) are an accurate measure of value of aspects of the course then we would expect a correspondence between them and the open ended "recall" item (Q 31). Further, if respondents are realistic in their assessment of what they will apply of what they have learned, we would expect them in Q 32 to commit to the quick and easy actions initially and not to list the more difficult skills and actions requiring the cooperation of others. This was directly tested by comparing their ratings and listings of valuable topics, and their listing of actions to be taken. Table 8 shows the rank ordering of mean scores from the management and AMTs samples, for the items in question 27,

Table 7
Product-moment validity coefficients: attitudes and indices, with 5 criteria (management $n = 163$/AMTs $n = 190$)

Training evaluation criteria	1 Communication	2 Assertiveness	3 Stress Management	4 Personal Styles Analysis	5 Problem solving/ role playing
Item number/name:					
1. We should avoid disagreeing with others		0.00/0.06			
2. We should avoid negative comments		0.11/0.12			
3. Conversation increases coordination	0.02/0.13				
4. Communication is as important as tech proficiency	0.05/0.06				
5. Sensitive to others problems	0.21**/0.18*			0.22**/0.18*	
6. Manager should take charge in emergency		− 0.15*/0.01			
7. Manager verbalizes plans	0.15*/0.000	0.13/0.000		0.16*/0.06	
8. Subordinates should not question manager's decisions		− 0.15*/ − 0.21**			
9. I perform well even when fatigued			− 0.05/0.06		
10. Managers should encourage questions	0.08/0.23**			0.17*/0.16*	
11. Subordinates should not take control anytime		− 0.06/0.01			
12. Debriefing/critique is important for coordination	0.21**/0.23**				0.06/0.21**
13. Technical proficiency equals successful mgt.				− 0.10/0.00	
14. Training is management's responsibility	0.00/0.08			0.07/0.06	
15. Team coordination is important in emergencies			0.06/0.11	23**/0.20**	
16. Pre-assignment briefings are effective					0.25**/0.20**
17. We should take other's personalities into account				0.22**/0.27**	
18. Managers should coordinate between work grps	0.14*/0.000				
19. Managers can leave personal problems behind			0.12/0.02		
20. My decision making is good even in crisis			0.01/0.10		
Index I: Command Responsibility		− 0.17*/0.01			
Index II: Communication and Coordination	0.28**/0.26**	0.23**/0.27**		0.31**/0.24**	
Index III: Avoiding Conflict		0.07/0.03			
Index IV: Recognition of Stressor Effects			− 0.02/ − 0.01		
Number of correlations	18	18	10	16	4
Percent statistically significant	50%	33%	0%	69%	75%

Table 8
Construct validity of criterion variables (Managment $n = 163$/AMTs $n = 190$)

1 Topics rated "Most value" (Criterion Var. Question 27)	2 Ranked means (Mgt./AMT)	3 Ranked frequency: good aspects of training (Question 31: Mgt./AMT)	4 Ranked frequency: will use on the job (Question 32: Mgt./AMT)
Active Listening[a]	— /1:4.42	—/6: Active Listening = 17	2: Listening = 39/3: Better Listening = 32
Assertiveness	1:5.54/4:4.25	1: Assertiveness = 28/4: Assertive = 21	4: Assertive = 19/5: Assertive = 27
Communication with others	2:4.43/2:4.31	3: Communication = 17/1:Comm. = 51	1: Communication = 43/1: Comm. = 55
Cases for Problem solving	3:4.42/3:4.29	6: Case studies = 11/5: Case studies = 20	8: Decision making = 7/12: Decision Mk = 4
Personal Style Analysis	4:4.28/5:4.12	2: Self-description Inv = 22/2:Self-desc. = 31	3: Self-awareness = 25/2: Self Aware = 36
Stress Management	6:4.20/7:4.10	4: Stress Mgt. = 15/3: Stress Mgt. = 27	9: Stress Mgt. = 5/10: Stress Mgt. = 12
Role Play for Prob. Solving	8:3.95/10:3.78	7: SubArctic Srvl. = 8/9: SubArctic Srvl. = 12	-/-
Norms[a]	—/8:4.09	5: Norms = 13/8: Norms = 13	7: Norms = 8/11: Norms = 9

[a]Not included in Question 27 for Management sample.

compared with the rankings of response frequencies for open-ended items 31 and 32. There is close agreement between the management and AMT samples for questions 27, 31, and 32 – with Spearman Rank Order Correlations of $\rho = 0.93$, 0.72, and 0.94 respectively. Table 8 reveals a close correspondence between what is rated as valuable (column 2) and what is subsequently recalled as "particularly good" (column 3) about the course. Rank Order Correlation shows the correspondence between question 27 and question 31 (Table 8 columns 2 and 3) to be $\rho = 0.75$ for the management sample and $\rho = 0.55$ for the AMTs (statistically significant beyond the 0.01 level).

5.3.1. *Expediency vs. value in the use of skills*

Examination of what will be used (column 4) shows that respondents, as expected, favor near term actions such as better use of listening skills, and a heightened awareness of self and others, instead of trying to develop an assertive style or making the effort to change work norms. This is demonstrated by many respondents stating that "communication", "listening" or "being aware of self and others" was their intended initial action. This can be explained as a preference for applying not what may be seen as most valuable (i.e., group problem solving or assertiveness), but what may be easiest to apply and what can be done by oneself – at least immediately following the training. Observing this expected effect provides evidence for construct validity. However the Rank Order Correlations for this pair of matched rankings show a higher degree of correspondence between column 2 (value) and column 4 (likely use) than that expediency *vs*. value explanation might predict – $\rho = 0.61$ for managers and $\rho = 0.85$ for AMTs (both significant beyond the 0.01 level of confidence).

5.4. *Is the MRM/TOQ valid*?

The management and AMT samples were in rather close agreement on their assessment both of the importance and usefulness of course topics. That concurrent validity data suggest that many of the individual attitude items correspond to the appropriate topics and course modules. Another test of concurrent validity (comparing multiple choice responses with answers written-in) confirms the "most valuable" course topics. The test for construct validity showed that those most valued topics were not always likely to be applied first

– they did not necessarily appear at the top of the list of behaviors most likely to be used immediately at the workplace.

6. Conclusion

The MRM/TOQ instrument shows integrity of factor structure and reasonable reliability. Although the tests of concurrent validity showed only two of four composite scales to relate to independent measures of course content, other analyses indicate that all the scales measure what they purport to measure. Measuring construct validity is important, but difficult. The various measures of recalled responses and recognized topics available in the questionnaire allowed a test of construct validity which indicated that the responses behaved as theoretically predicted – thus lending credence to the meaning of the concepts together as a construct of usable resource management skills. Overall, these results demonstrate the respondents' ability to understand both the open ended and the multiple response scales and their motivation to respond to them thoughtfully. The MRM/TOQ has reasonable psychometric characteristics, and it can be used as a good measure to evaluate communication and MRM training programs in other aviation maintenance settings.

Acknowledgements

This work was accomplished with the assistance of Socio-technical Design Consultants, Inc., in cooperation with Institute of Safety and Systems Management, University of Southern California. Dr. B. Ann Bettencourt and Dr. Seong Woo-Choi, research assistants during 1991–92 and 1994–95 respectively, performed many of the statistical analyses supporting the findings reported here. The research was sponsored initially by the Biomedical and Behavioral Sciences Branch of the FAA Office of Aviation Medicine and later by the Office of Life and Microgravity Sciences and Applications of the National Aeronautics and Space Administration (Contract NCC2-812). The present analysis and report was prepared with funding under Federal Aviation Administration Research Grant #96-G-003.

References

Cronbach, L.J., 1951. Coefficient Alpha and the internal structure of tests. Psychometrika 16, 297–334.

Cronbach, L.J., Meehl, P.E., 1955. Construct validity in psychological tests. Psychological Bulletin 52, 281–302.

Gregorich, S.E., Helmreich, R.L., Wilhelm, J.A., 1990. The structure of cockpit management attitudes. Journal of Applied Psychology 75, 682–690.

Guilford, J.P., 1936. Fundamental Statistics in Psychology and Education. McGraw-Hill, New York.

Helmreich, R.L., Foushee, H.C., Benson, R., Russini, R., 1986. Cockpit management attitudes: exploring the attitude-performance linkage. Aviation, Space and Environmental Medicine 57, 1198–1200.

Norusis, M.J., 1990. SPSS/PC + Statistics 4.0. Chicago, Illinois.

Selltiz, C., Wrightsman, L.S., Cook, S.W., 1976. Research Methods in Social Relations. Holt, Rinehart and Winston, New York.

Taggart, W., 1990. Introducing CRM into maintenance training. Proceedings of the Third International Symposium on Human Factors in Aircraft Maintenance and Inspection. Federal Aviation Administration, Washington DC.

Taylor, J.C., Robertson, M.M., 1995. The effects of crew resource management (CRM) training in airline maintenance: results following three years experience. NASA Contractor Report 196696/Contract NCC2-812. Washington, D.C., National Aeronautics and Space Administration.

Taylor, J.C., Robertson, M.M., Choi, S., 1997. Empirical results of maintenance resource management training for aviation maintenance technicians. Proceedings of the Ninth International Symposium on Aviation Psychology. The Ohio State University, Columbus, OH.

Taylor, J.C., 2000. The development and evolution of maintenance resource management (MRM). International Journal of Industrial Ergonomics (Special Issue on Human Factors in Aviation Maintenance), in press.